Electrophysiology of statistical learning: exploring the online learning process and offline learning product

Mikhail Ordin[1,2,*], Leona Polyanskaya[1], David Soto[1,2], Nicola Molinaro[1,2]

[1] BCBL – Basque Centre on Cognition, Brain and Language, Donostia, Spain

[1,2] IKERBASQUE – Basque Foundation for Science, Bilbao, Spain

[*]Corresponding author:
Mikhail Ordin
m.ordin@bcbl.eu
Basque Centre for Cognition, Brain and Language (BCBL)
Mikeletegi 69, Donostia, 20009, Spain

KEYWORDS: statistical learning, speech segmentation, transitional probabilities

RUNNING TITLE: *Statistics and memory in speech segmentation*

ABSTRACT:

A continuous stream of syllables is segmented into discrete constituents based on the transitional probabilities (TPs) between adjacent syllables by means of statistical learning. However, we still do not know whether people attend to high TPs between frequently co-occurring syllables and cluster them together as parts of the discrete constituents or attend to low TPs aligned with the edges between the constituents and extract them as whole units. Earlier studies on TP-based segmentation also have not distinguished between the segmentation process (how people segment continuous speech) and the learning product (what is learnt by means of statistical learning mechanisms). In the current study we explored the learning outcome separately from the learning process, focusing on three possible learning products: holistic constituents that are retrieved from memory during the recognition test, clusters of frequently co-occurring syllables, or a set of statistical regularities which can be used to reconstruct legitimate candidates for discrete constituents during the recognition test. Our data suggests that people employ boundary-finding mechanisms during online segmentation by attending to low inter-syllabic TPs during familiarization and also identify potential candidates for discrete constituents based on their statistical congruency with rules extracted during the learning process. Memory representations of recurrent constituents embedded in the continuous speech stream during familiarization facilitate subsequent recognition of these discrete constituents.

1 INTRODUCTION

Segmenting the continuously changing world into units is a daunting yet unavoidable challenge the brain faces as it transforms ongoing, dynamic sensory flow into structured and meaningful experience. Continuous speech, for example, must be divided into sentences, phrases and words by means of statistical learning mechanisms (SL). SL is not restricted to processing only linguistic input (Abla & Okanoya, 2008). To a large extent, segmentation relies on tracking transitional probabilities (TPs) between smaller elements (e.g., syllables, distinct tones or pseudo-shapes). Speech-like acoustic streams with high TPs between syllables are easily extracted and committed to memory during habituation learning (e.g., Aslin et al., 1998). During subsequent recognition tests, these sequences can be identified as legitimate word candidates and mapped to objects more readily than their random counterparts (Graf Estes et al., 2007; Hay et al., 2011).

Despite the fact that TPs are known to be crucial for SL, the cognitive mechanisms for TP-based segmentation are still under debate. Two possible classes of mechanisms, suggesting different roles for TPs in the segmentation of continuous input, have been proposed: 1) boundary-finding mechanisms and 2) clustering mechanisms (see Perruchet & Pacton, 2006 for an overview). Boundary-finding mechanisms rely on detecting the edges of the constituents by attending to the troughs in TPs, which are aligned with the boundaries between holistic consecutive segments. (Endress & Mehler, 2009; Elman, 1990). By contrast, clustering mechanisms rely on attention to TPs between syllables that have a higher probability of co-occurrence, leading to the emergence of clusters that do not necessarily correspond to whole constituents (Perruchet & Poulin-Charronnat, 2012; Perruchet, P., & Vinter, 1998; Frank, Goldwater, Griffiths & Tenenbaum, 2010). We also do not have a clear understanding of what is learnt by means of SL. There are three possible alternatives for the learning product: (i) recurrent syllable sequences as whole discrete constituents; (ii) frequently presented chunks, e.g., frequently co-occurring syllable pairs; (iii) statistical regularities per se, e.g., TPs.

In this study, we wanted to find out (1) whether people attend to lower or to higher TPs for the purpose of online segmentation, i.e., during exposure to continuous sensory input, and (2) what is

endorsed as a legitimate constituent offline, i.e., during the recognition test following the familiarization exposure. The former question is related to the process of segmentation, that is, how people use TPs to extract discrete constituents from continuous flow, and the latter is related to the product, that is, what is actually learnt by means of SL mechanisms.

To address these issues, we set up an experiment using the artificial language learning paradigm (Saffran et al., 1996), and constructed an artificial language that could lead to the emergence of false word candidates that had never appeared in the habituation acoustic stream - *phantoms* (Endress & Mehler, 2009; Endress & Langus, 2017). Imagine a stream of syllables, in which syllable A is followed by syllable B, and the syllabic pair AB is frequently presented in the habituation input as a part of the syllable triplet ABC. However, syllable B can also be followed by syllable Z, and the syllabic pair BZ is frequently presented as a part of a tri-syllabic YBZ sequence. Although the tri-syllabic sequence ABZ never occurs during learning, it could potentially be endorsed as a sequence present in the habituation input during the recognition test (Endress & Mehler, 2009; Turk-Browne & Scholl, 2009; Endress & Langus, 2017). The constructed artificial language stream for our study included 12 recurrent syllable triplets (statistical *words*), and the TPs between adjacent syllables allowed for the emergence of phantoms as perceptual units (see the *Methods* section for details regarding the experimental material). The habituation phase was followed by a dual forced-choice recognition test. For this test, we paired *words* with *phantoms* (triplets that shared the statistical properties of the words but had never occurred during habituation and thus could not have been encoded and committed to memory as whole constituents) and with *non-words* (random tri-syllabic sequences). On each trial, participants were asked to give confidence ratings regarding their choice. We measured neural oscillatory changes related to the encoding phase, and the behavioral and ERP responses related to recognition performance. Neural changes during the habituation phase were related to the learning process and to extracting regularities for building representations (Battering & Paller, 2017; Ding et al., 2016; 2017a). Online segmentation was indexed by the entrainment of neural oscillations to the frequency of extracted constituents; increases in power peaks and precision (inter-trial coherence) were considered to be a function

of learning progress. During the subsequent recognition phase, the behavioral and ERP responses to words, phantoms and non-words informed us about what is learnt by means of SL mechanisms.

1.1 Electrophysiology of TP-based segmentation

The learning process and the role of TPs in segmentation can be explored by tracking ongoing changes in neural oscillatory dynamics as habituation progresses. When we process real connected speech, multiple neural oscillations track extracted linguistic units at different timescales by tuning into syllabic, phrasal and sentence rhythms: power peaks and coherence maxima in the EEG signal emerge at frequencies that correspond to the rate of these linguistic constituents (Peelle & Davis, 2012; Peelle, Gross, & Davis, 2013; Ghitza, Giraud, & Poeppel, 2013; Ding et al., 2016; 2017a). The adjustment of brain rhythms to the phase and period of rhythms in environmental stimuli is known as neural entrainment. The spectrum of the temporal envelope of sound, which reflects how fast sound intensity fluctuates over time, reveals well-separated power peaks at around 4-5Hz across typologically different languages, corresponding to the mean syllabic rate across languages (Ding et al., 2017b). This modulation spectrum is a purely acoustic measure of syllabic rhythm, and thus provides unambiguous acoustic cues for brain-to-sound coupling. Rhythms at higher levels of the linguistic hierarchy are less consistent across languages in terms of their acoustic manifestation, and sometimes totally non-existent. Nevertheless, neural oscillations also emerge at the frequency of higher-level linguistic constituents. Online encoding of statistical words during familiarization with an artificial language is indexed by the entrainment of brain rhythm to word frequency, and the degree of entrainment increases as the learning progresses (Batterink & Paller, 2017; Buiatti et al., 2009). The degree of neural entrainment correlates with the percentage of endorsed statistical words during the subsequent recognition test. We assume that the degree of neural entrainment will correlate with the percentage of recognized phantoms only if segmentation is supported by clustering mechanisms, and frequently co-occurring syllable pairs represent the learning outcome of statistical learning. If the learning product is instead represented by whole triplets, because segmentation relies on boundary-finding mechanisms, the degree of neural entrainment should correlate with the percentage of endorsed triplets during the recognition test.

Ongoing changes in oscillatory dynamics during familiarization will inform us about the segmentation process. Neural oscillations at the syllabic frequency are expected to emerge immediately as exposure to the acoustic stream starts. The entrainment of cortical rhythms to the acoustic syllabic rhythm is a physiological response of the brain to the acoustic structure of the ambient signal (Greenberg & Ainsworth, 2004). It does not reflect higher-level processing and is not expected to be modulated over the course of habituation. Neural oscillations at word-frequency are expected to emerge only when tri-syllabic word boundaries have been identified and should increase in power and phase coherence with exposure (Wöstmann, Fiedler, & Oblesser, 2017). If TPs are used to detect syllable transitions with low probabilities and to insert boundaries between tri-syllabic constituents, we should observe the emergence and gradual increase of neural oscillations at one third of the syllable frequency rate. If, on the other hand, TPs are used to detect syllables that frequently cluster together, then we should expect the emergence of neural oscillations at the frequency of boundaries between frequently co-occurring syllable pairs. As the syllable pairs that cluster together are overlap (syllables 1 and 2 and syllables 2 and 3 constitute frequent syllable clusters), we should observe an increase in neural oscillations at the frequency of the syllable as exposure increases.

In addition to behavioural data, the product of statistical learning will be explored by examining ERPs elicited by triplets, phantoms and foils during the recognition test. If only whole tri-syllabic constituents are encoded as learning products, then ERPs elicited by words are going to be different from those elicited by phantoms and by foils. If statistical rules are used to endorse legitimate word candidates, then the ERPs elicited by non-words should be different from the ERPs elicited by statistically congruent tokens. Finally, if phantoms are accepted as legitimate constituents because they can be reconstructed from frequent syllable pairs, while words are recognized upon their retrieval from memory as whole units, then endorsing words and phantoms would rely on different neural substrates (Skosnik et al., 2002), and thus may have distinct electrophysiological correlates (ERPs) even in the absence of different behavioural responses to words and phantoms.

2 MATERIAL AND METHODS

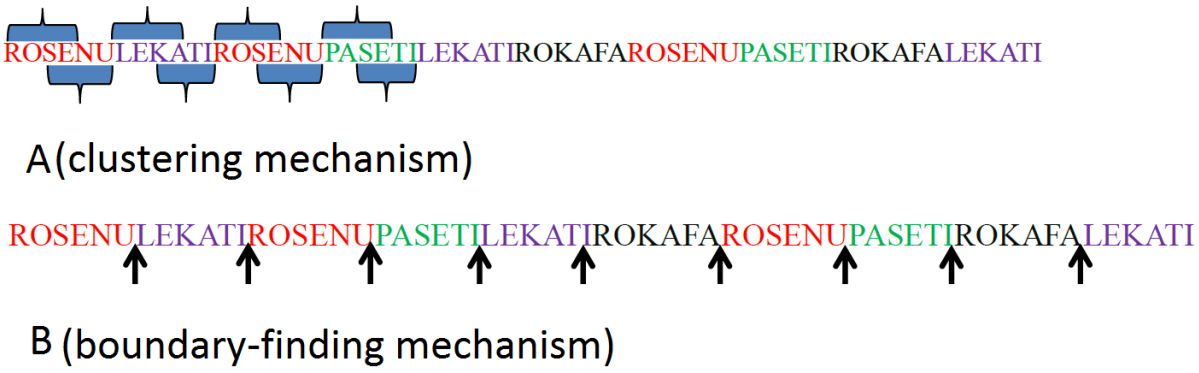
2.1 Participants

34 Spanish-Basque bilinguals between 18 and 26 years of age (with no known history of neurological diseases or hearing problems, right-handed) were recruited and received monetary compensation for their time. The age of acquisition for their second language (Basque) was three years (at school), with Spanish being the first language and the family language of all participants. This is the most common linguistic profile for participants in the Basque country, where the experiment was carried out. The study was approved by the BCBL Ethics committee (approval number 280317). All participants signed a written informed consent form prior to the study.

2.2 Stimuli

The syllable inventory included 18 syllables (consonant + vowel) that were used to make 12 tri-syllabic statistical nonsense words, each syllable contributing to two words. Each syllable was 240ms in duration (vowels–140ms). Each statistical word had a unique combination of consonants. We also counterbalanced the position of plosive, sonorant and fricative consonants. Each consonant was used an equal number of times. We tried to counterbalance the vowels and the position of vowels within statistical words. The statistical words were randomly concatenated 125 times into a continuous stream (1080sec), ramped on both ends to prevent the listener from using the edge syllables of streams as anchors for the edges of the initial and final statistical words. In the final acoustic stream, the TPs between syllables within words were 50%, and the TPs between syllables straddling the statistical word boundaries were around 15%. The list of words is presented in **Table 1**. **Figure 1** presents an example of a stream, demonstrating the different potential roles of TPs in segmentation. The acoustic stream was created using MBROLA speech synthesizer using ES1 voice, the stimuli were prepared at 22050Hz. F0 was set to 120Hz. The stream was preceded by 293.76sec of silence, so that we could use the EEG signal measured during the resting state period as a baseline.

INSERT FIGURE 1 SOMEWHERE HERE¨



The TPs implemented in the habituation stream allowed for the emergence of *phantoms* as perceptual units (tri-syllabic sequences that did not appear in the habituation acoustic stream as whole triplets, yet had 50% TPs between syllables and were composed of the syllable pairs that constituted the words).

For the recognition test, each word and phantom was synthesized as a separate file. Finally, we prepared random tri-syllabic sequences with inter-syllable TPs set to 0% (i.e., sequences composed of syllabic pairs that never appeared in the habituation stream, we refer to these sequences as *non-words* below). None of the words, phantoms or non-words resembled any real word from the participants' native languages. The list of phantoms and non-words is in **Table 1**.

INSERT TABLE 1 SOMEWHERE HERE

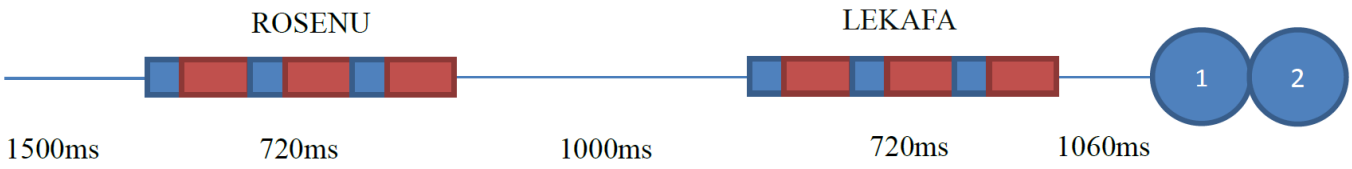
2.3 Experimental procedure

The experiment consisted of two separate phases. During the familiarization phase, participants were explicitly instructed to detect and memorize the words of an unknown "extraterrestrial" language. Immediately after the learning phase, segmentation performance was assessed by means of a dual forced-choice recognition test. During the test, each word was used in 4 different pairs, pitted against two different non-words and two different phantoms, counterbalancing the position of items in the pairs. Also, each phantom was pitted against two different non-words, counterbalancing the position of items in the pair. In total, 72 test pairs were created. Each item was used an equal number of times in the test. The acoustic stimuli were presented auditorily, and while the stimuli were being played, the fixation cross was shown in the middle of the screen. Items within the test pairs were separated by 1000ms. Both the habituation stream

and test stimuli were presented in PsychoPy (v.1.80.04) via two speakers positioned about one meter in front of the participants, at approximately 75dB. Presentation of each pair was followed by two questions. First, participants were asked to choose whether the first or the second item in each pair was the word from the extraterrestrial language (i.e., recognition response). To test whether participants were more aware of the words than the phantoms, we asked them to indicate how confident they were in their answer (i.e., confidence response). The confidence rating assigned to each trial was used as an awareness index of the behavioral responses (Schwiedrzik, Singer, & Melloni, 2011). Confidence ratings were given on a 4-point scale: 1-guess, 2-slightly confident, 3-fairly confident, 4-sure. We encouraged participants to use the full range of the scale.

The participants were instructed not to move/blink when the cross was on the screen, and to input their response only when the screen disappeared and the prompt to give a response was presented on the screen. The prompt to give the recognition response was given 1060ms after the presentation of the second item in the stimulus had finished. The structure of the test trial is presented in **Figure 2**.

INSERT FIGURE 2 SOMEWHERE HERE



2.4 EEG data acquisition and pre-processing

EEG recordings were recorded throughout the experiment, both during the habituation and testing phases using the Brain Amp DC acquisition system. EEG was recorded from 27 Ag/AgCL electrodes (Fp1/2, F7/8, F3/4, FC5/6, FC1/2, T7/8, C3/4, CP5/6, CP1/2, P7/8, P3/4, O1/2, Fz, Cz, Pz) positioned according to the 10/20 system positioned in an EasyCap and referenced to the left mastoid. The sampling rate was set at 500 Hz. Additional electrodes were placed on the right mastoid, at the outer canthi of both eyes, and above and below the right eyes. Impedances were kept below 5kΩ.

The EEG signal recorded during the testing phase was re-referenced offline to the average of the left and right mastoids and low-pass filtered (30Hz, 24dB/octave, using a 50Hz notch filter to remove electrical interference). This continuous EEG signal was checked for the presence of artifacts such as EMG bursts, glitches, spikes and other instrumental artifacts, artifacts caused by movements, etc. Artifacts were defined as segments +/- 200ms around amplitude changes exceeding 75µV/ms, exceeding a 100µV difference over a 50ms segment, or when the amplitude was lower than 0.5µV over a 50ms segment.

For the ocular correction, the EEG signal was decomposed into independent components using a fast ICA extended algorithm implemented in BrainVision Analyzer 2.0.2.5859 based on the entire dataset for each participant. Components that captured blinks, horizontal eye movements and, in rare cases, heartbeats (identified by visual inspection based on the components' topography and energy) were removed. The mean number of removed components was 2.063.

2.5 Oscillatory analysis, encoding

Since we were interested in evaluating possible hemispheric differences with EEG (as reported in Vanvooren et al., 2014), the recordings during the Encoding phase were first referenced to Cz, which has been reported to be the best reference for this purpose (van der Reijden et al., 2005; Van Dun et al., 2009; Vanvooren et al., 2014). No reliable hemispheric difference was found, in line with Ding et al. (2017). Raw data were then segmented into epochs of 8.64 seconds corresponding to the presentation of 12 tri-syllabic words in the continuous stream. This time window gives a frequency resolution for the Fourier transform of each trial equal to 0.12Hz, and allows good estimations at the frequencies of interest (1.39 Hz, 2.08 Hz, and 4.17 Hz). We then grouped the first 40 repetitions in Section 1, the next 40 in Section 2 and the third 40 in Section 3. For baseline purposes, we also considered 40 consecutive 8.64-second intervals extracted from the resting state period collected before the main experiment began. We then averaged the epochs in each Section and computed the inter-trial phase coherence (ITC) across trials for each Section: we first computed the phase of the Fourier-transformed signal in the frequency domain (using the function *angle* in Matlab R2012b), and then computed phase coherence across trials. ITC provides a measure of spectral consistency across trials, without showing the typically skewed distribution of power estimates. However, since these

frequency tagging effects should also emerge in power (Ding et al., 2017), we also computed the absolute value of the discrete Fourier transform (DFT) of the data as implemented in Matlab R2012b. This reflects the power of the EEG responses synchronized to the auditory stimulus. The resulting ITC and DFT estimations were further used for grand-averaging across participants.

Power and coherence at the frequencies of interest were contrasted statistically considering all electrodes. We also exploratively tried to group electrodes in different clusters, but different clustering did not yield any differences in the result pattern, and the effect was always significant across all electrodes. We first evaluated if the peak at each frequency was statistically stronger for the stimulation sections compared to the resting state and then conducted pairwise contrasts for the three Sections. Our main hypothesis was that these parameters should become more positive during exposure across the different Sections. Consequently, we tested the right tail of the statistical distribution.

2.5 ERP analysis, recognition test

The continuous EEG signal was segmented into 1420ms epochs, yielding 48 epochs corresponding to each token type (phantoms, words and non-words). The epochs were time-locked to the onset of each item in each test pair, with 200ms preceding the onset and 1220ms following the onset of the first syllables in the items. The epochs that contained artifacts were rejected. We excluded participants from further analysis if we had to reject over 25% of trials in any of the conditions (words, phantoms or non-words). In total, 4 participants were excluded from both the ERP and oscillatory analyses. For the remaining 30 participants, the average proportion of trials that survived artefact rejection was 95.5%, comprising 93.6% for words, 96.4% for phantoms, and 96.3% for non-words. Epochs were baseline corrected (subtraction method) using the 200ms of data preceding the onset of each item.

As we did not have an a priori hypothesis regarding the time windows in which ERP amplitudes should differ between the conditions, we decided to follow a data-driven approach to determine the latencies of interest. For this purpose, differences in amplitude between conditions were assessed by means of a non-parametric cluster-based permutation statistical approach implemented in the Fieldtrip toolbox. This

algorithm evaluates the presence of significant effects in the data by employing a cluster-based approach while controlling for multiple comparisons (Maris & Oostenveld, 2007). In the interval between 0 and 1 sec, for each sample of the data the null hypothesis that the data in the experimental conditions come from the same probability distribution is tested. The null distribution was extracted by randomly permuting the values of the contrasted conditions 1000 times and selecting the largest cluster-level statistic in each permutation. Cluster-level statistics with an F-value for the three conditions highlighted a significant effect in the time-window on which the pairwise comparisons were made using the same non-parametric cluster-based permutation statistical approach. Clusters falling in the highest or lowest 2.5^{th} percentile of the permutation distribution were considered significant.

3 RESULTS

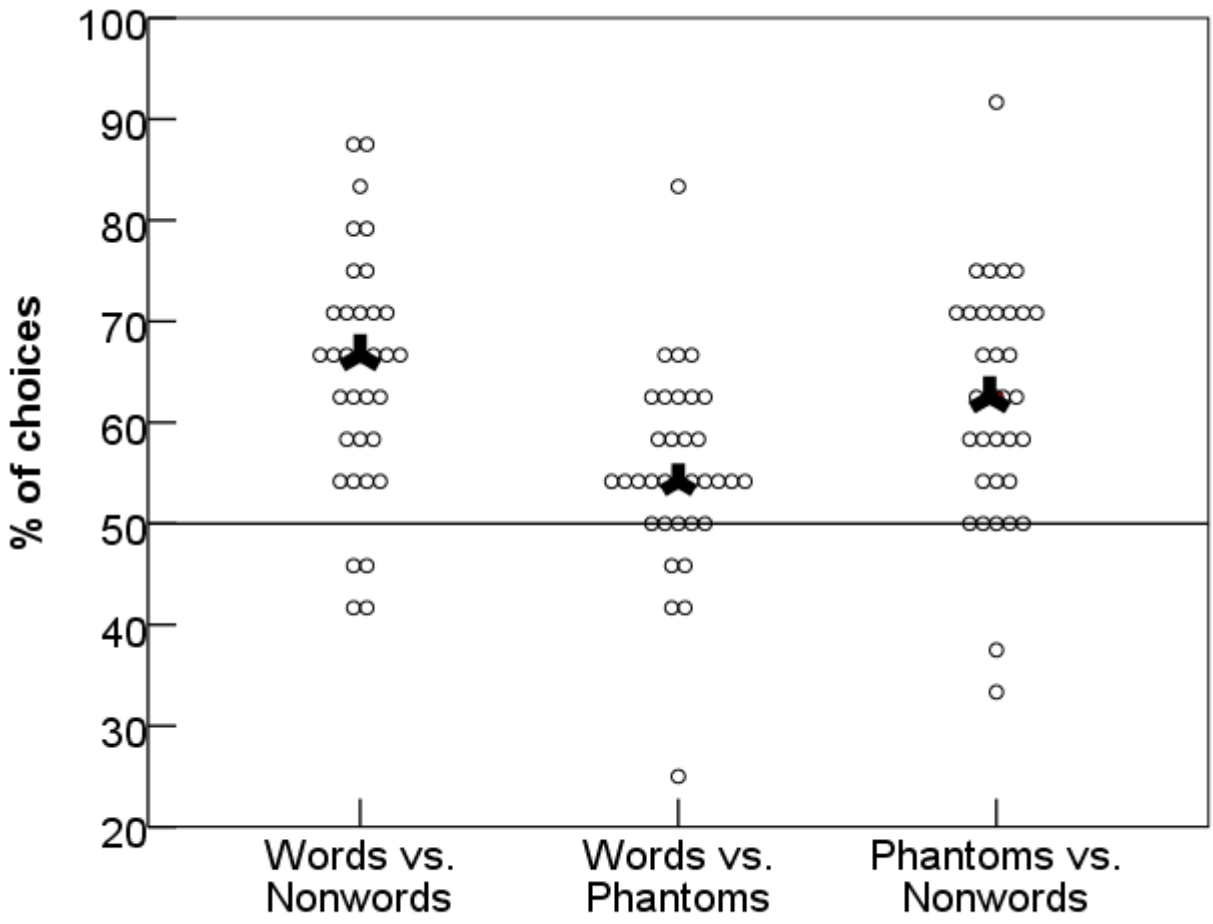
3.1 Behavioural data

As shown in **Figure 3**, participants preferred words over non-words significantly more often than would be expected by chance (M=64.83% [10.71, 18.94], 95% confidence interval in square brackets, SD=11.79, t(33)=7.331, p< .0005, Cohen's d=1.26). Participants also preferred words to phantoms more often than would be expected by chance (M=55.15% [1.77, 8.52], SD=9.68, t(33)=3.1, p=.004, Cohen's d=.53). When choosing between phantoms and non-words, participants preferred phantoms (M=62.13% [8.03, 16.23], SD=11.76, t(33)=6.017, p< .0005, Cohen's d=1.03). The results show that participants tended to reject non-words as discrete constituents of the continuous habituation acoustic stream. The magnitude of the effect shows that the non-words were rejected equally efficiently when paired with words (d=1.26) and with phantoms (d=1.03), while the magnitude of the effect was substantially lower for test trials in which the participants had to choose between words and phantoms (d=.53). The preference for words against phantoms is significantly weaker than the preference for words against non-words (M=-9.68 [-13.94, -5.42], SD=12.21, t(33)=-4.623, p< .0005, Cohen's d=.79) and the preference of phantoms against non-words (M=-6.99 [-11.9, -2.07], SD=14.09, t(33)=-2.89, p=.007, Cohen's d=.64). However, we did not observe a

significant difference between preference of words against non-words and preference of phantoms against non-words (M=2.7 [-1.48, 6.87], SD=11.96, t(33)=1.315, p=.198, Cohen's d=.23).

These results suggest that phantoms are endorsed as word candidates when paired with non-words, but not when paired with statistical words from the habituation stream. Non-words were rejected with equal accuracy, irrespective of whether they were paired with words or phantoms.

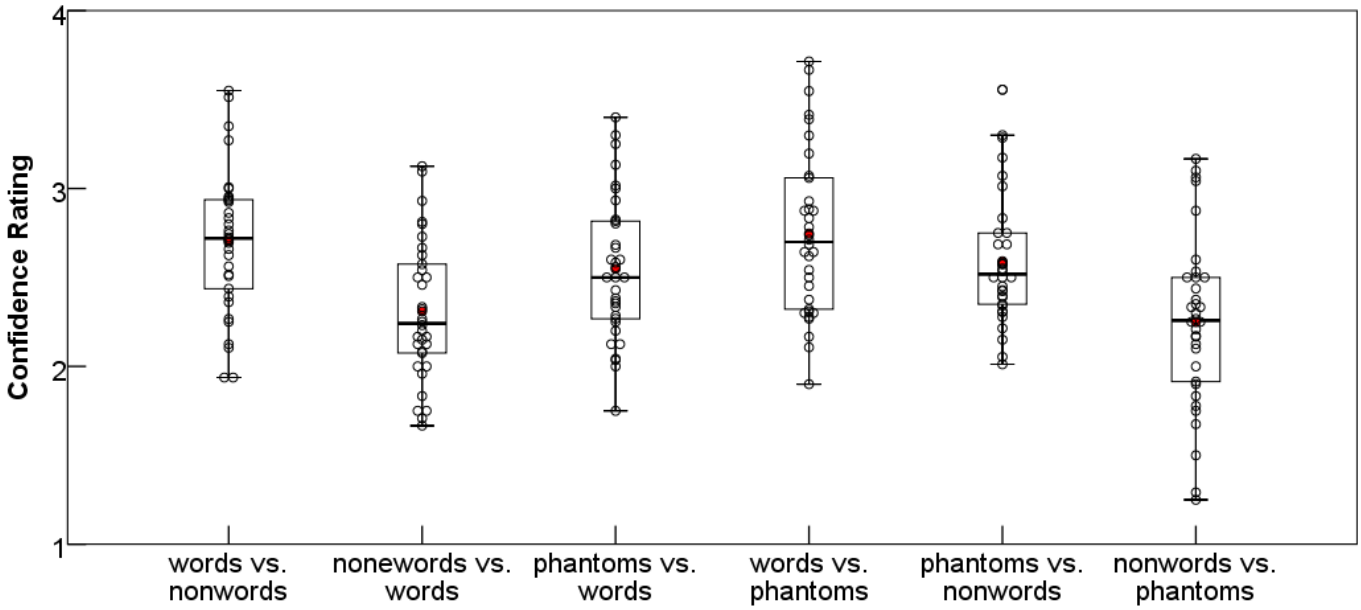
INSERT FIGURE 3 SOMEWHERE HERE



As shown in **Figure 4**, participants assigned the lowest confidence ratings to trials in which they chose non-words, while the highest confidence ratings were assigned to trials in which participants chose words. The difference between confidence ratings assigned to responses in which participants chose words (M=2.72), phantoms (M=2.57), or non-words (M=2.28) was significant ($F(2,66)=23.073$, $p< .0005$, $\eta_p^2=.411$, sphericity assumed based on Mauchly's test, p=.142). Pairwise comparisons (Bonferroni correction applied,

all the tests performed on weighted averages of confidence ratings) showed that opting for words elicited significantly higher confidence than opting for phantoms (M=.144[.033, .254], SD=.316, t(33)=2.652, p=.024, Cohen's d=.455), and that opting for phantoms elicited significantly higher confidence than opting for non-words (M=.291 [.141, .441], SD=.429, t(33)=3.956, p< .0005, Cohen's d=.678). Confidence ratings assigned to correct responses tend to be higher than confidence ratings assigned to incorrect responses on trials when participants are consciously aware of the learning product (Galvin et al., 2003). The pattern or results in our study suggest that people consider words to be more legitimate than phantoms, and this also aligns with their recognition accuracy.

INSERT FIGURE 4 SOMEWHERE HERE



3.2 EEG: Learning phase

These analyses were performed to reveal the dynamics of neural oscillations during the habituation phase. We aimed to understand whether brain rhythms emerge at the frequency of syllables, frequent syllable pairs, or recurrent syllable triplets, and whether these oscillations are modulated by the length of exposure. The increase in power and inter-trial coherence of brain oscillations with exposure reflects the learning process and shows whether the continuous signal is split into recurrent syllable sequences or frequent syllable pairs.

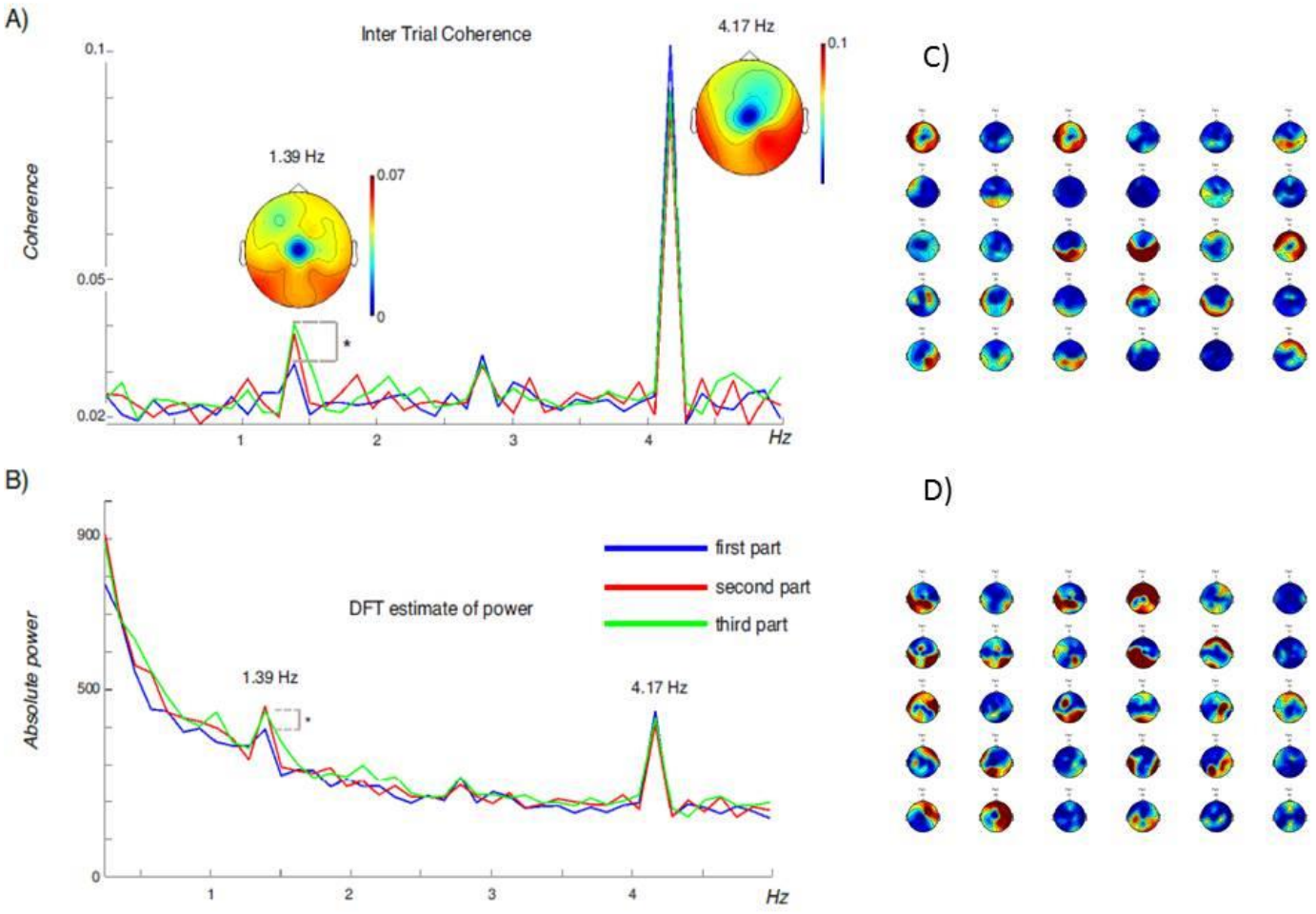
These oscillatory dynamics reveal whether TPs function to mark the edges of the holistic constituents (low inter-syllable TPs), or to raise the perceptual salience of frequent syllable clusters (high inter-syllable TPs).

The ITC (inter-trial coherence) analyses showed two main peaks of entrainment compared to resting state recordings (**Figure 5**). Syllable-level (4.17 Hz: $t(29) = 6.31$, $p<0.001$) and word-level (1.39 Hz: $t(29) = 2.05$, $p=0.025$) peaks were evident across all electrodes (see Figure 6 for individual topoplots). No peak was evident at 2.08 Hz. The harmonic of the word-level peak was evident at 2.78 Hz. We did not detect any difference in syllable-level peaks between sections (all t statistics <1). When considering the word-level peak, the comparison between Section 1 and Section 2 revealed a significant difference at $p=0.08$, $t(29) = 1.44$, the comparison between Section 1 and Section 3 revealed a significant difference at $p=0.027$, $t(29) = 2.01$. No significant difference emerged between Section 2 and Section 3, $t(29) = 0.41$, $p=.68$.

In the DFT analysis (power spectrum analysis) of the EEG signal recorded during the habituation phase, the syllable-level peak at 4.17 Hz ($t(29) = 7.18$, $p<0.001$) and the word-level peak at 1.39 Hz ($t(29) = 3.67$, $p<0.01$) showed larger amplitudes compared to the resting condition (**Figure 5**). Again, at 2.08 Hz we did not observe any effect ($t(29)<0.5$). In the syllable-level peak at 4.17 Hz no reliable differences emerged between the three Sections (all t statistics < 1). A power peak evident at 2.78 Hz compared to the resting state condition reflected a harmonic of the word-level peak, again, with no statistically significant differences in amplitude between the three Sections (all t statistics <.1). In the word-level peak at 1.39 Hz an eloquent trend emerged when comparing Section 1 and Section 2 ($t(29) = 1.61$, $p=0.059$), and a statistically significant effect emerged from the comparison between Section 1 and Section 3 ($t(29) = 1.85$, $p=0.037$). No meaningful or statistically significant effect emerged from the Section 2-Section 3 comparison ($t(29) = 0.43$).

INSERT FIGURE 5 SOMEWHERE HERE

INSERT FIGURE 6 SOMEWHERE HERE

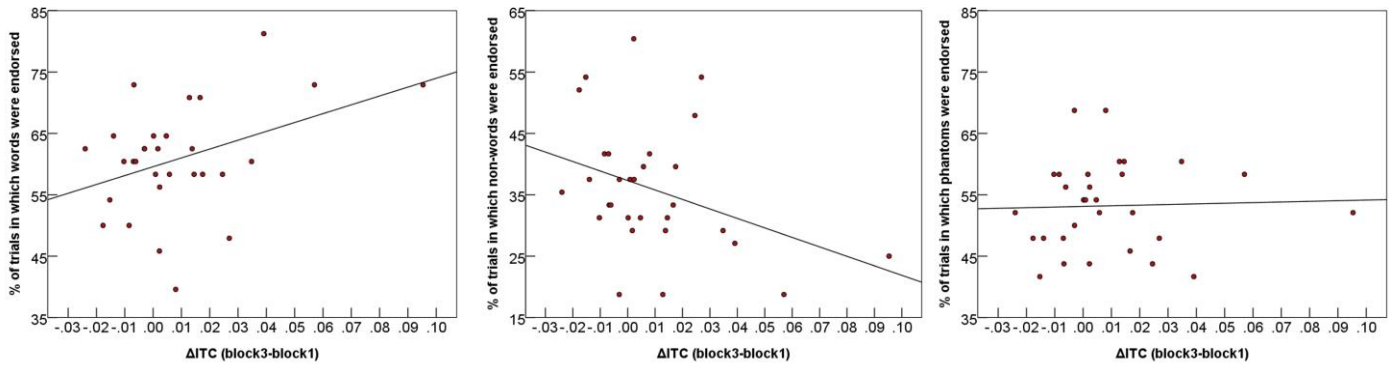


In order to evaluate the specificity of the oscillatory activity at 1.39Hz, we contrasted both power and coherence at this frequency averaged across exposure segments (first, second and third segments of the exposure periods) and channels. We then averaged the activity in the neighbor frequencies (1.27Hz and 1.5Hz) and contrasted those values (across participants) against our target frequency with a permutation approach employing 50,000 reiterations. We observed a significant effect at 1.39Hz compared to the neighbor frequencies for both power ($t=5.25$, $p<0.001$) and inter-trial coherence ($t=4.17$, $p<0.001$). This indicated that the expected word-level effect was statistically reliable compared to the rest of the oscillatory activity.

We further investigated the relationships between the Encoding and Recognition phases by testing for correlations between the frequency results and recognition performance. We subtracted the estimates of Section 1 (assumed to reflect initial exposure) from the estimates of Section 3 (supposed to reflect acquired

word-level knowledge) for each participant (averaged across all electrodes). We tested for correlations between differences in entrainment estimates (ΔITC) and recognition performance measured as the percentage of trials in which words, phantoms or non-words were endorsed. Scatterplots illustrating these correlations are displayed in **Figure 6.** We did not detect any outliers in measures of neural entrainment or behavioural performance (the variance did not exceed 3SD around the mean, or exceed the value 2.58, when the measures were z-transformed, for any variable). However, due to the very different scale of ITC gain, on the one hand, and the percentage scores, on the other hand, the problem of heteroscedasticity in the data is inevitable. Therefore, we used a non-parametric Spearman´s method for correlational analysis. We observed a positive but not significant correlation between the percentage of trials in which participants selected a word from the pair of test items (when one of the items was indeed a word) and the ΔITC, rho = 0.151 [-.224, .484], p=.257). Importantly, we observed a negative relation, also not significant, between the percentage of trials in which participants selected a non-word (when one of the items was indeed a non-word) and the ΔITC, rho = – .239 [-.551, .132], p=.07. The correlation between the ΔITC and the preference of participants for phantoms was negligibly weak, rho= .074 [-.294, .422], p=.577. We later compared the strength of these correlations pairwise, and found no significant differences, z=1.433, p=.126 for ΔITC and percentage of trials with endorsed words vs. ΔITC and percentage of trials with endorsed non-words; z=.282, p=.389 for ΔITC and percentage of trials with endorsed words vs. ΔITC and percentage of trials with endorsed phantoms; and z=.979, p=.164 for ΔITC and percentage of trials with endorsed non-words vs. ΔITC and percentage of trials with endorsed phantoms. These correlations suggest that the increase in ITC at the word frequency is linked with increased recognition accuracy for words and for non-words, and the weakest association is between the proportion of endorsed phantoms and ITC gain. This tentatively indicates that it is unlikely that phantoms emerge as perceptual units during learning. However, the correlations are not significant, and we did not detect significant differences between the strengths of correlations between ΔITC and performance measures, indicating that any conclusion regarding the association between the entrainment and behavioural variables are only suggestive, and the data does not yield convincing evidence against the null hypothesis. The correlations should be interpreted with caution and re-evaluated against a larger sample size.

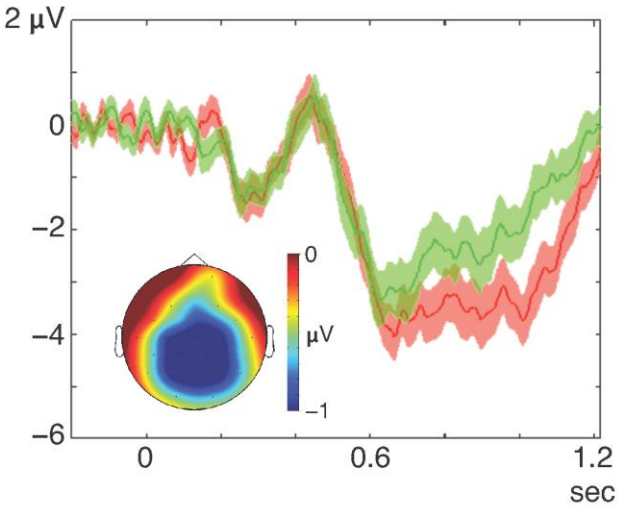
INSERT FIGURE 7 SOMEWHERE HERE

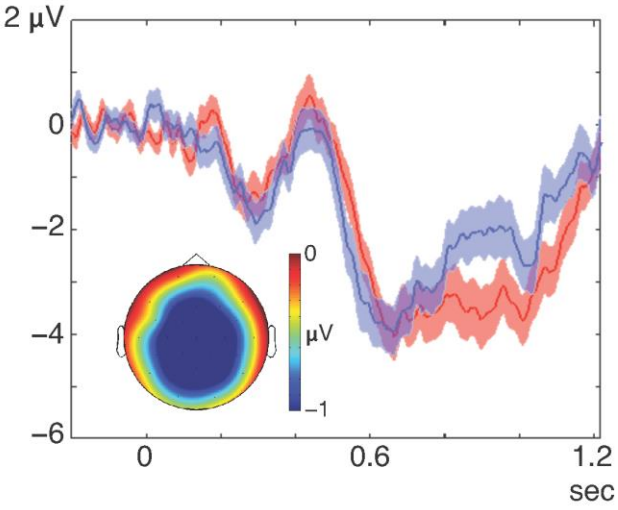


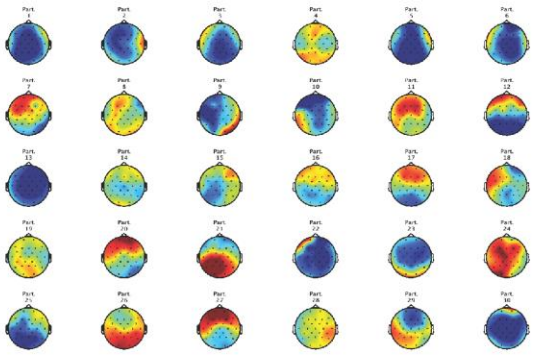
3.3 ERP: Recognition test

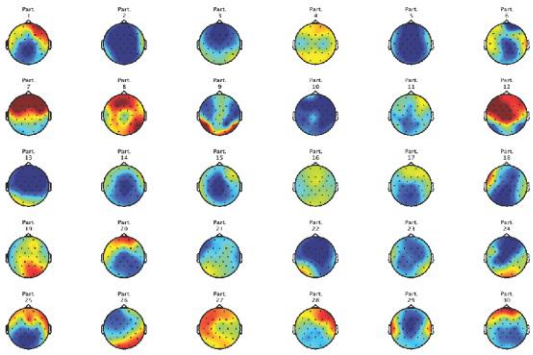
The comparison between ERPs elicited by words, phantoms and non-words showed a significant difference ($p<.001$) in the time interval between 836ms and 952ms following the onset of the stimulus (the time interval was determined based on a data-driven, non-parametric cluster-based permutation statistical approach (Maris & Oostenveld, 2007), in the interval between 0 and 1sec following stimulus onset, see the Methods section for details). Pairwise comparison confirmed that non-words elicited more negativity compared to both words ($p=<.05$) and phantoms ($p=<.01$). These negative effects showed a central-posterior distribution over the scalp (**Figure 8**). No statistically significant effect emerged from the comparison between words and phantoms. To estimate the degree of support for the null hypothesis in the words vs. phantoms comparison, we averaged the ERP amplitude over the central electrodes (C3, C4, P3, P4, Cz, Pz, CP1, CP2) during the time interval between 836ms and 952ms for words and phantoms, and calculated the Bayes Factors (BF) using the Bayesian paired t-test (performed in JASP). $BF_{10}=0.371$ showed that the null model is favoured 2.7 times more than the alternative model, which provides only weak evidence (Raftery, 1995) for a lack of difference in ERP amplitudes between words and phantoms, and calls for further investigation of this issue.

INSERT FIGURE 8 SOMEWHERE HERE


2 µV
0
−2
−4
−6
0
0.6
1.2
sec
0
µV
−1


2 µV
0
−2
−4
−6
0
0.6
1.2
sec
0
µV
−1





4 DISCUSSION

The analysis of the EEG signal recorded during habituation revealed entrainment of neural oscillations to the acoustic signal at the syllable frequency rate. However, brain rhythm at the syllable frequency rate was not modulated by length of exposure, and thus the entrainment at this frequency likely represents a physiological response to the environmental stimuli. In contrast, the emergence of neural oscillations at the word frequency rate increased both in power and in ITC as a function of exposure length, reflecting the learning process. ITC and power measures provide supplementary, not overlapping information regarding neural entrainment (Wöstmann, Fiedler, & Oblesser, 2017). ITC informs us about the precision of entrainment, while power tells us about the strength of entrainment. Entrainment can potentially be strong but not necessarily precise. We did not observe any evidence that frequent syllabic clusters emerged as perceptual units during familiarization, neither in regard to the strength, nor precision of neural entrainemnt. This pattern of results suggests that holistic statistical recurrent words, not frequent clusters of syllables are extracted online from continuous speech. Statistical cues are used to detect syllable pairs with low TPs and to insert boundaries between recurrent triplets in continuous acoustic signal.

The results of correlational analysis indicate at possible association of the ITC gain during learning and recognition accuracy during the test for correctly endorsed words. Correlations between the gain in precision of neural oscillations and the proportion of trials in which phantoms were endorsed as word candidates is negligible (rho=.074), providing tentative support for the earlier interpretation that words, not phantoms emerged as perceptual units during familiarization. However, the difference in the strength of correlation between ΔITC with the percentage of accepted words and accepted phantoms was not significant, suggesting that the presumption of no difference is not yet overcome by the data, and another study with a larger sample size would be needed to establish whether the difference in correlation strength is genuine.

During recognition performance, words were differentiated from both phantoms and from non-words, but with different levels of accuracy and confidence. Recognition was better for words relative to non-words and phantoms, while the difference between words and phantoms, although significant, was substantially

weaker. As we did not find any supporting evidence for a clustering mechanism during segmentation, we suggest that phantoms are endorsed based on the statistical congruency that they share with words. Both words and phantoms can be reconstructed based on transitional probabilities during the test phase, which explains why phantoms are sometimes confused with words. This also suggests that an artificial grammar (i.e., a network of TPs) is one of the learning outputs of the statistical learning mechanism.

Choosing between two structured syllabic sequences, both statistically congruent with the rules of the artificial language (i.e., words and phantoms), is more challenging than choosing between statistically congruent (grammatical) and statistically incongruent (agrammatical) sequences. We suggest that people detect this absence of statistical structure and distinguish between grammatical and agrammatical test tokens based on TPs. When words and phantoms are pitted against each other, both test items exhibit similar internal statistical structure, and the need to recourse to additional cues for making a decision makes the task more difficult. This suggestion is also supported by the analysis of ERPs. Ungrammatical tokens (i.e., nonwords) elicited a different ERP than grammatical sequences (words and phantoms), with more negativity spread over the central-parietal scalp area. Considering that ERPs may be taken to reflect the similarity between two cognitive processes (Kutas & Federmeier, 2011), we tentatively propose that the recognition of statistically congruent items relies on similar cognitive processes, irrespective of whether they were actually presented during habituation. This suggests that both words and phantoms are endorsed based on their statistical congruency with the rules embedded in the familiarization stream, providing support for the hypothesis that statistical regularity per se is a primary product of statistical learning. It is important to point out, however, that the Bayesian analysis only weakly supported a lack of difference between the electrophysiological response to words and phantoms, indicating that this issue warrants further investigation.

It is not rare to have inconsistent neurophysiological and behavioural results; however, neurophysiological results are usually more sensitive to experimental manipulations than behavioural results. In this study we observed the reverse pattern: ERPs elicited by words and phantoms did not differ significantly, while at the behavioural level, words were preferred to phantoms, and trials in which words had been endorsed were

assigned higher confidence ratings than trials in which phantoms were endorsed over non-words. This deviation from the general tendency to detect finer distinctions in neurophysiological data may have a variety of explanations. The number of trials per condition might not have been sufficient to achieve a good signal-to-noise ratio for differentiating between electrophysiological responses to words and phantoms, which may differ only slightly from each other, in contrast to the larger differences between statistically congruent (words and phantoms) vs. incongruent (non-words) items. However, individual variability within conditions **(Figure 7)** is lower than variability between conditions, suggesting that the signal-to-noise ratio is sufficiently high to detect the existing effect, and the lack of differences in ERPs elicited by words and phantoms cannot be explained by low statistical power and type II error. Another possible explanation for the discrepancy between the sensitivity of the behavioural and electrophysiological data is the elusive nature of ERP responses. Different neurophysiological processes can be manifested in similar ERP components. For example, memory processes, recognition processes, lexico-semantic integration, phonological and orthographical processing and a range of other cognitive processes may be manifested in similar N400-like ERP components (see Kutas & Federmeier, 2011 for a review). Statistical learning and the recognition of statistically extracted tokens engages a wide network of neural substrates, suggesting highly distributed neural generators (Paraskevopoulos et al., 2017), which may function in a dynamic interaction. Statistical learning is mediated by intraparietal areas, pre-motor areas in inferior frontal cortices, and temporal areas, which are activated dynamically (Paraskevopoulos et al., 2017). Further, different ensembles may produce similar neural signatures, yet result in different behaviours (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017). A third explanation is based on a dual-process model of recognition memory (Rugg & Curran, 2007), which suggests that recognition is based on distinct processes eliciting a familiarity effect followed by a recollection effect, which occurs in a later time-window after the onset of the event to be recognized. The detected difference in ERP responses to words and phantoms vs. non-words may represent the combined effects of detecting deviations from statistical regularities acquired during the learning phase and familiarity elicited by statistical congruency. There would be a longer latency for the stronger mnemonic effects related to recollection, and thus differentiating between words presented during the learning stage and phantoms, could be beyond the analyzable time window in our study due to design

limitations. As participants were allowed to make their behavioural decisions later, the effect of recollection, which is not present in neurophysiological data at an earlier latency, might facilitate recognition on trials with words, leading to a higher confidence rating when endorsing words and a slight preference for words over phantoms. Further investigation is required to ascertain which of these explanations best accounts for the pattern of results reported here.

The timing and topography of the increased negativity elicited by non-words, compared to the negativity elicited by words and phantoms, allows us to define this effect as belonging to the family of N400s. Here, the N400 is considered to be a heuristic definition of stimulus-related brain activity around 400ms-600ms after the critical event, eliciting *relative* negativity on deviations from an (over)learnt pattern (Kutas & Federmeier, 2011). In the auditory modality, the N400 exhibits slightly more frontal and less right-biased (more central) scalp distribution, which corresponds to the topography of the observed negativity effect. In our study, the critical event that leads to the N400-like effect is the onset of the second syllable, when the listener detects a violation of the internal statistical structure of the syllabic sequence. A non-word is a deviation from a pattern of syllable sequencing that the participant has learnt while listening to the familiarization stream. This leads us to suggest that the increased negativity we found might be the electrophysiological correlate of detecting a structural violation.

Our results are also in line with earlier studies on N400-like effects in recognition memory (Friedman & Johnson, 2000; Curran, 2000): new items elicit more negativity than old, (over)learnt items. N400 is elicited by unlearnt items when the participant is not necessarily aware of the specific details of the (over)learnt items that allowed recognizing them as old (Smith & Guster, 1993; Curran, 2000). Danker et al. (2008) found that negativity associated with new verbal stimuli is shifted forward and to the centre of the scalp, compared to the typical N400 elicited in language research studies (thus the term frontal, or fN400). This also corresponds to the topography of negativity distribution we observed in our study in response to the non-words. The fN400 is limited to situations in which the stimuli are not processed holistically, but when categorization of stimuli into old and new requires unitization, i.e., using the attributes of stimuli to integrate them into familiar (old) and unfamiliar (new) units (Mecklinger, 1998). These data suggest a different

interpretation for the obtained pattern of results. It is possible that statistically congruent items – words and phantoms – trigger a familiarity effect, because both are reconstructed based on the artificial grammar that the participants acquired during familiarization. Reconstruction of triplets from syllables for further categorization of test items as grammatical (familiar) or agrammatical creates a situation of unitization that leads to an fN400 associated with new items (non-words). Interestingly, we explicitly asked the participants to detect and memorize the words of a novel language during the habituation phase. We did not inform them that the embedded words had any kind of internal structure. Learning the artificial grammar, as opposed to extracting the recurrent triplets, happened without intention. Rugg and Curran (2007) claimed that the ERP index of familiarity actually reflects implicit learning. Both words and phantoms elicit similar degrees of familiarity (and thus similar electrophysiological differences from nonwords), and distinguishing between them requires explicit retrieval of words as holistic constituents from memory. The data suggest that statistical congruency has more weight for recognition of legitimate structural constituents, while memory representations facilitate recognition of triplets that were actually embedded in the continuous signal and presented multiple times during learning. The facilitatory effect of memory representations for holistic constituents was also reflected in the higher confidence ratings assigned to endorsed words than to endorsed phantoms.

Vos and Paller (2007) and Pallet et al. (2007) argued that the processes responsible for eliciting the N400 are also active during recognition memory and recognition tasks, and thus may have the same or overlapping neural generators, which include lateral pre-frontal and temporal areas. The neural substrates in these areas also support statistical learning processes and recognition of newly learnt sequences. Paraskevopoulos et al. (2017) showed that statistical learning relies on a distributed network composed of the superior temporal gyri (bilateral), where TPs are calculated (Bischoff-Grethe et al., 2000; McNealy et al., 2006), bilateral intra-parietal areas, which support memory processes (Ciaramelli et al., 2008; Vilberg & Rugg, 2008), and the left inferior frontal gyrus (Karuza et al., 2013), where phonological processes are mapped onto motor codes, with the arcuate fasciculus supporting functional and structural connectivity between temporal and pre-motor areas for integration of motor-auditory information (López-Barroso et al., 2013). This audio-pre-

motor interface supports the initial learning and recognition of new phonological words (Rodríguez-Fornells et al., 2009; López-Barroso et al., 2013).

Overall, the current results show that entrainment of neural oscillations to acoustic rhythm at the syllable frequency is automatic (e.g. happening incidentally, without conscious intention, Ding et al., 2017) and to a large extent is determined by the physiology of the peripheral auditory system (Greenberg & Ainsworth, 2004). Entrainment of neural oscillations to the acoustic signal at the word frequency is a marker of the segmentation progress (Battering & Paller, 2017; Ding et al., 2016; 2017; Buiatti et al., 2009). It therefore also reflects the learning process which is modulated by the length of exposure. ITC of neural oscillations at the word frequency reflects the precision of neural entrainment and predicts recognition accuracy only for words. These results suggest that the role of the TPs in online segmentation is to detect (or predict) the boundaries between recurrent structural constituents, i.e., syllabic triplets. These triplets, as well as the network of transitional probabilities (i.e., artificial grammar) are the target of the statistical learning mechanisms.

5 CONCLUSION

A continuous stream of syllables is segmented into discrete constituents based on transitional probabilities (TPs) between adjacent syllables by means of statistical learning (SL). Here we focused on: 1) the role of TPs in segmentation; 2) what is learnt by means of SL. Listeners could use TPs to detect frequent clusters with high inter-syllabic TPs, or to detect syllable pairs with low TPs at the edges of holistic recurrent constituents. We addressed these hypotheses within an artificial language learning paradigm by measuring changes in the EEG oscillatory dynamics during learning, and by measuring behavioral and ERP responses during a subsequent recognition test. The results indicate that participants segmented recurrent triplets, not frequent clusters of syllables, during habituation. During the recognition test, both words (old triplets embedded in the familiarization stream) and phantoms (novel yet statistically congruent triplets) tended to be endorsed as legitimate word candidates. However, old triplets were recognized better and with higher confidence than novel ones. Random sequences were not recognized. Random, statistically incongruent

sequences elicited ERPs that differed from those elicited by statistically congruent old and new tokens. This indicates that both statistical regularities and whole constituents segmented from continuous sensory input are learnt by means of SL mechanisms. TPs, however, have more weight in the recognition of segmented constituents, while memory representations facilitate recognition of old triplets.

The present study highlights the need to focus on the mechanisms that allow humans to successfully process a continuous flow of information with an unlimited degree of complexity using limited cognitive resources. One of these mechanisms is statistical learning. Statistical learning engages a set of general cognitive mechanisms that are used to extract regularities and discrete constituents from continuous sensory input. Although these mechanisms did not evolve specifically for speech processing, they are brought to bear when we listen to and learn a real language and should be further explored to better understand speech and language processing.

ACKNOWLEDGEMENTS

The research was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) through the "Severo Ochoa" Programme for Centres/Units of Excellence in R&D (SEV-2015-490), project grants RTI2018-098317-B-I00 to MO and a Basque Government grant PI-2017-25 to DS. LP was supported by the European Commission through the Marie Skłodowska-Curie Research Fellowship.

CONFLICT OF INTEREST

Authors declare no conflict of interest

ABBREVIATIONS

CI, confidence interval; DFT, discrete Fourier transform; EEG, electroencephalography; ERP, event-related potential; ITC, inter-trial phase coherence; M, mean difference; SD, standard deviation; Δ, difference.

AUTHORS' CONTRIBUTION

MO and LP conceived the project and ran the experiment, MO, LP and NM analyzed the data, all authors discussed the results and wrote the manuscript.

DATA ACCESSIBILITY

Raw data, and analysis data, experiment and analysis scripts are available from the first (corresponding) author upon request (mordin@bcbl.eu). Also, we have deposited the raw and pre-processed data on Figshare repository (DOI: 10.6084/m9.figshare.9989000 for the EEG data recorded during the familiarization phase and DOI: 10.6084/m9.figshare.9988610 for the EEG recordings made during the recognition test (also available as segmented into epochs per condition, with bad trials rejected).

REFERENCES

Aslin, R., Saffran, J., & Newport, E. (1998). Computation of conditional probability statistics. Psychological Science 9, 321-324.

Batterink, L., & Paller, K. (2017). Online neural monitoring of statistical learning. Cortex, 90, 31-45.

Bischoff-Grethe, A., Proper, S. M., Mao, H., Daniels, K. A., & Berns, G. S. (2000). Conscious and unconscious processing of nonverbal predictability in Wernicke's area. *Journal of Neuroscience 20*, 1975–1981.

Buiatti, M., Peña, M., & Dehaene-Lambertz, G. (2009). Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. Neuroimage 44(2), 509-519.

Ciaramelli, E., Grady, C.L., & Moscovitch M. (2008). Top-down and bottom-up attention to memory: a hypothesis (AtoM) on the role of the posterior parietal cortex in memory retrieval. *Neuropsychologia 46(7)*, 1828–51.

Curran, T. (2000). Brain potentials of recollection and familiarity. Memory and Cognition 28, 923–938.

Danker, J., Hwang, G., Gauthier, L., Geller, A., Kahana, M., & Sekulera, R. (2008). Characterizing the ERP Old–New effect in a short-term memory task. Psychophysiology 45(5), 784-793.

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017a). Characterizing Neural Entrainment to Hierarchical Linguistic Units using Electroencephalography (EEG). Frontiers in Human Neuroscience 11: 481.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. Nature Neuroscience 19, 158–164.

Ding, N., Patel, A., Chen, L., Butler, H., Luo, C., Poeppel, D. (2017b). Temporal modulations in speech and music. Neuroscience & Biobehavioral Reviews 81, 181-187.

Elman, J. L. (1990). Finding structure in time. Cognitive Science 14, 179–211.

Endress, A. & Langus, A. (2017). Transitional probabilities count more than frequency, but might not be used for memorization. Cognitive Psychology, 92, 37-64.

Endress, A., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. Journal of Memory and Language, 60(3), 351-367.

Frank, M. C., Goldwater, S., Griffiths, T., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. Cognition 117, 107–125.

Friedman, D., Johson, R.E. (2000). Event-Related Potential (ERP) studies of memory encoding and retrieval: A selective review. Microscopy Research and Technique 51, 6–28.

Galvin, S. J., Podd, J. V., Drga, V., and Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. Psychonomic Bulletin and Review 10, 843–876.

Ghitza, O., Giraud, A.-L., & Poeppel, D. (2013). Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence. *Frontiers in Human Neuroscience* 6:340.

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. Psychological Science 18, 254–260.

Greenberg, S., & Ainsworth, W. (2004). Speech processing in the auditory system: An Overview. In S. Greenberg, W. Ainsworth, A. Popper, and R. Fay (Eds.). Speech Processing in the Auditory System (pp. 1-62). New York, NY: Springer-Verlag.

Hay, J., Pelucchi, B., Graf Estes, K., & Saffran, J. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. Cognitive Psychology 63, 93-106.

Karuza, E. A., Newport, E. L., Aslin, R. N., Starling, S. J., Tivarus, M. E., & Bavelier, D. (2013). The neural correlates of statistical learning in a word segmentation task: An fMRI study. *Brain and Language 127*, 46–54.

Krakauer, J., Ghazanfar, A., Gomez-Marin, A., MacIver, M., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron 93(3)*, 480-490.

Kutas, M., & Federmeier, K. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). The Annual Review of Psychology 62, 621-647.

López-Barroso, D., Catani, M., Ripollés, P., Dell'Acqua, F., Rodríguez-Fornells, A., & de Diego-Balaguer, R. (2013). Word learning is mediated by the left arcuate fasciculus. *Proceedings of the National Academy of Sciences 110(32)*, 13168-13173.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods 164(1), 177–190.

McNealy, K., Mazziotta, J. C., & Dapretto, M. (2006). Cracking the language code: Neural mechanisms underlying speech parsing. *Journal of Neuroscience 26*, 7629–7639.

Mecklinger, A. (1998). On the modularity of recognition memory for object form and spatial location: A topographic ERP analysis. Neuropsychologia, 36(5), 441-460.

Orban, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. Proceedings of the National Academy of Sciences of the United States of America, 105(7), 2745–2750.

Paller, K.A., Voss, J.L., & Boehm, S.G. (2007). Validating neural correlates of familiarity. *Trends in Cognitive Sciences 11(6)*, 243–250.

Paraskevopoulos, E., Chalas, N., & Bamidis, P. (2017). Functional connectivity of the cortical network supporting statistical learning in musicians and nonmusicians: an MEG study. *Scientific reports 7(1)*, 16268.

Peelle, J. E., and Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. Front. Lang. Sci. 3:320.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cerebral Cortex 23(6), 1378-87.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. Trends in Cognitive Sciences 10, 233–238.

Perruchet, P., & Poulin-Charronnat, B. (2012). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. Journal of Memory and Language, 66, 807-818.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. Journal of Memory and Language 39, 246–263.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), Sociological methodology (pp. 111–196). Cambridge, MA: Blackwell.

Rodríguez-Fornells, A., Cunillera, T., Mestres-Missé, A., & de Diego-Balaguer, R. (2009). Neurophysiological mechanisms involved in language learning in adults. *Philosophical Transactions of the Royal Society B: Biological Sciences 364(1536)*, 3711–3735.

Rugg, M., & Curran, T. (2007). Event-relation potentials and recognition memory. Trends in Cognitive Science 11(6), 251-257.

Saffran, J., Johnson, E., Aslin, R., &Newport, E. (1999). Statistical learning of tone sequences by human adults and infants. Cognition, 70, 27-52.

Schwiedrzik, C., Singer, W., & Melloni, L. (2011). Subjective and objective learning effects dissociate in space and in time. PNAS 108, 4506–4511.

Skosnik, P.D., Mirza, F., Gitelman, D.R., Parrish, T.B., Mesulam, M.M., & Reber, P.J. (2002). Neural correlates of artificial grammar learning. Neuroimage 17(3), 1306-1314.

Smith, M.E., Guster, K. (1993). Decomposition of recognition memory event-related potentials yields target, repetition, and retrieval effects. Electroencephalography and Clinical Neurophysiology 86, 335–43.

Turk-Browne, N. & Scholl, B. (2009). Flexible visual statistical learning: Transfer across space and time. Journal of Experimental Psychology: Human Perception and Performance, 35, 195–20.

van der Reijden CS, Mens LH, Snik AF (2005) EEG derivations providing auditory steady-state responses with high signal-to-noise ratios in infants. Ear Hear 26:299–309.

van Dun B, Wouters J, Moonen M (2009) Optimal electrode selection for multi-channel electroencephalogram based detection of auditory steady-state responses. J Acoust Soc Am 126:254–268.

Vanvooren, S., Poelmans, H., Hofmann, M., Ghesquière, P., & Wouters, J. (2014). Hemispheric asymmetry in auditory processing of speech envelope modulations in prereading children. Journal of Neuroscience 34(4), 1523-1529.

Vilberg, K.L., & Rugg, M.D. (2008). Memory retrieval and the parietal cortex: a review of evidence from a dual-process perspective. *Neuropsychologia 46(7)*:1787–99.

Voss, J.L., & Paller, K.A. (2007). Neural correlates of conceptual implicit memory and their contamination of putative neural correlates of explicit memory. *Learning and Memory 14(4)*, 259–267.

Wöstmann, M., Fiedler, L., & Obleser, J. (2017). Tracking the signal, cracking the code: speech and speech comprehension in non-invasive human electrophysiology. Language, Cognition and Neuroscience 32(7), 855-869.

TABLES

Table 1. The list of words, phantoms and non-words used in the experiment

Words	**Phantoms**	**Nonwords**
ROSENU	PASENU	ROTIMO
ROKAFA	LEKAFA	SEPAKO
PASETI	ROSETI	FALUSA
LEKATI	ROKATI	FOLERI
PAMONU	PAMOFA	TAMUPE
LEMOFA	LEMONU	NIKANU
PERIKO	MURIKO	NURIPE
MURIFO	LUTASA	FOLUKA
PETASA	PERIFO	NIMUKO
LUTAFO	PETAFO	MOPARO
MUNIKO	MUNISA	LESATI
LUNISA	LUNIKO	TASEFA

FIGURE CAPTIONS

Figure 1. Cognitive mechanisms of TP-based segmentation (TP-transitional probability). **A.** Representation of clustering mechanism. TPs can be used to extract clusters of syllables (in this case, frequent syllable pairs) that can be recombined to recognize the discrete structural constituents of the continuous stream. If the "clustering" segmentation mechanism is engaged, *phantoms* can emerge as perceptual units and be recognized as discrete structural constituents of the continuous signal. For example, the syllable pairs ROSE and SENU compose the recurrent triplet ROSENU. The pairs PASE and SETI constitute. The triplet PASETI. From these statistical words, a triplet ROSETI can be reconstructed, with the same TPs between syllables Yet with no memory traces of ROSETI as a holistic unit because it is never implemented as a whole triplet in familiarization stream. **B.** Representation of boundary-finding mechanism. TPs can be used to detect the syllable pairs with low transitional probability between syllables, and to "insert" boundaries corresponding to the locations of the troughs in the stream of TPs, "bracketing" the recurrent triplets and extracting them as holistic discrete constituents.

Figure 2. The structure of the test stimuli pairs. Each item in the pair is 720ms in duration (vowels are 140ms, consonants are 100ms. The participant listens to a pair of stimuli, and then he needs to decide whether the first or the second syllabic sequence in the test pair is a word from the language he listened to. Afterwards, the participant has to indicate how sure he is in his response, on a 4-point scale. "1" in the circle stands for recognition response, "2" stands for confidence response.

Figure 3. Recognition results. Preference for words when paired with non-words (left) and phantoms (middle), and preference for phantoms when paired with non-words (right). The points indicate the percentage of words, phantoms and non-words chosen in each comparison for each individual.

Figure 4. Confidence ratings. Confidence ratings assigned to words, phantoms and non-words when these items are paired versus possible alternatives. The chosen item is always the first one in the name of the test pair on the abscise axis. That is, "words vs. non-words" denotes all the pairs in which words were paired with non-words, irrespective of whether the word was the first or the second item in the test trial, and in which the participants endorsed "words". "non-words vs. words" denotes all the pairs in which words were paired with non-words, irrespective of whether the word was the first or the second item in the test trial, and in which the participants endorsed "non-words".

Figure 5. Neural entrainment during learning. **A)** Inter-trial phase coherence as a function of the learning stream Section (blue – section 1, red – section 2, and green – section 3). Averaged topographical plots showing the distribution of ITC across the scalp at the frequency of words (1.39Hz) and syllables (4.17Hz). **B)** Power spectrum as a function of the learning stream Section.

Figure 6. Individual topoplots. A) Individual topoplots showing the distribution of ITC across the scalp at the frequency of syllables. B) Individual topoplots showing the distribution of ITC across the scalp at the frequency of words.

Figure 7. Correlations between inter-trial phase coherence (ITC) gain during learning and proportion of endorsed tokens during recognition test. **Left:** Correlations between differential ITC (difference in the ITC between the first and the third exposure blocks) and the overall preference of participants for non-words over phantoms and words. **Middle:** Correlations between differential ITC and the overall preference of participants for phantoms over words and nonwords. **Right:** Correlations between differential ITC and the overall preference of participants for words over phantoms and non-words.

Figure 8. ERPs elicited by nonwords (red) and words (green), on the left, and by nonwords (red) and phantoms (green), on the right. Shaded areas show two standard errors (30 participants) around the mean at each timepoint. Topoplots (averaged topoplots inside the ERP graphs and individual topoplots below) show the distribution of the effects, computed by subtracting the mean amplitude of words (left) or phantoms (right) from the mean amplitude of non-words in the time windows that yielded significant differences between three conditions in the non-parametric cluster-based permutation analysis (836-952ms).